## *A.* Comparison with other graph-based models

The graph structure learning algorithm used in this study is an improvement of the graph neural network (GNN) algorithm, which can effectively learn the differences CDA graph topology distribution and optimize GNN parameters. To verify whether this strategy is helpful in predicting CDAs, we conducted a comparison experiment between these two models. Additionally, we also added comparisons to the graph convolutional network (GCN) model to increase persuasiveness. To be specific, we adopt the idea of ablation experiments, keeping the overall framework of the GSLCDA model constant and replacing only the features used in this model with those extracted by the GNN and GCN algorithm. The 5CV experimental results obtained by the GNN and GCN models on the benchmark dataset are listed in Table S1 and its ROC curves are displayed in Figures S1 and S2. The comparison of the values shows that the GNN model is 2.72% less accurate than the GSLCDA model, and 2.42%, 0.62%, 5.10% and 0.0259 lower on F1, sensitivity, MCC and AUC, respectively. The GCN model is also 1.47%, 1.55%, 3.13%, 2.93% and 0.0189 lower than GSLCDA on these evaluation parameters. These experimental outcomes suggest that the graph structure learning algorithm used in this study can more deeply mine CDA data features than GNN and GCN algorithm, which contributes more to the GSLCDA performance in predicting potential CDAs.

TABLE S1. 5CV results obtained by other graph-based model on CircR2Disease dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fold | Accu. (%) | F1 (%) | Sen. (%) | MCC (%) | AUC |
| 1st fold | 90.50 | 90.69 | 92.56 | 81.06 | 0.8908 |
| 2nd fold | 90.50 | 90.84 | 97.44 | 81.87 | 0.9075 |
| 3rd fold | 88.84 | 90.18 | 93.23 | 77.52 | 0.8654 |
| 4th fold | 89.26 | 89.68 | 96.58 | 79.47 | 0.9131 |
| 5th fold | 90.65 | 90.91 | 96.64 | 81.96 | 0.9332 |
| |  | | --- | | GNN Model | | **89.95±0.84** | **90.46±0.52** | **95.29±2.22** | **80.38±1.88** | **0.9020±0.0255** |
| 1st fold | 91.86 | 91.49 | 88.97 | 83.83 | 0.9093 |
| 2nd fold | 91.19 | 91.22 | 93.10 | 82.45 | 0.9054 |
| 3rd fold | 90.85 | 90.97 | 92.52 | 81.74 | 0.9124 |
| 4th fold | 90.17 | 90.43 | 91.95 | 80.38 | 0.8980 |
| 5th fold | 91.95 | 92.55 | 97.39 | 84.33 | 0.9200 |
| |  | | --- | | GCN Model | | **91.20±0.74** | **91.33±0.78** | **92.78±3.03** | **82.55±1.60** | **0.9090±0.0081** |
| GSLCDA | **92.67±0.51** | **92.88±0.71** | **95.91±0.59** | **85.48±0.97** | **0.9279±0.0165** |



Fig. S1. ROC curves obtained by GNN model on CircR2Disease dataset

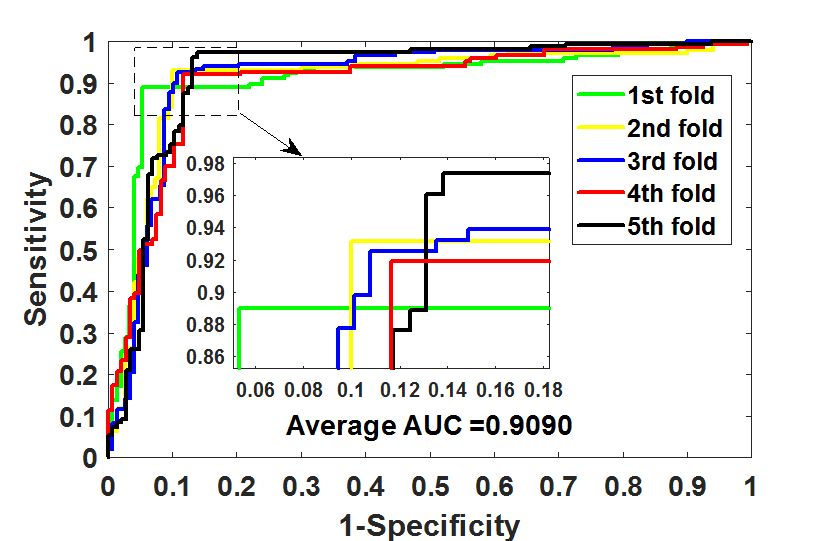


Fig. S2. ROC curves obtained by GCN model on CircR2Disease dataset

## *B.* Four benchmark datasets for circRNA-disease

we evaluated the GSLCDA on four circRNA-disease datasets, which are CircR2Disease, CircAtlas, Circ2Disease and CircRNADisease. These four datasets were obtained by different researchers by collecting and organizing experimental data on circRNA and disease related literature. We conducted an overall evaluation of the comprehensive performance of the GSLCDA model using the CircR2Dissease dataset in the experiment, and evaluated the generalization and universality of the GSLCDA model using other three datasets CircAtlas, Circ2Disease, and CircRNADissease. Due to different collection methods, times and literature searches, the samples and data they contain are also different. After removing redundancy and noise, the specific sample data information used for the experiment is summarized in table S2. More details about these four datasets are as follows:

CircR2Disease is a dedicated database and comprehensive platform for disease-related circRNAs. The experimentally supported database host 739 entries that including 725 circRNA-disease associations, 661 circRNAs, 100 diseases by curating from published literatures. The cured information could be applied to investigate the function and molecular mechanisms of circRNAs in disease. Users could freely view it for providing a timely and valuable resource for circRNAs research clinical application. CircR2Disease can be accessed at http://bioinfo.snnu.edu.cn/CircR2Disease/index.aspx

CircAtlas database integrates a collection of circulating transcripts and is based on 1070 RNA-seq samples collected from 19 normal tissues of 6 vertebrates. This database contains 1,007,087 highly reliable circRNAs, of which over 81.3% have been assembled into full-length sequences. The developer profile their expression pattern, conservation, and functional annotation. And the developer describe a novel multiple conservation score, co-expression, and regulatory networks for circRNA annotation and prioritization. CircAtlas can be accessed at https://ngdc.cncb.ac.cn/circatlas/index.php

Circ2Disease is a database that intergrated a comprehensive collection of human circRNA and disease associations. All those associations were manually retriewed from public literature. The developer also predicted the corresponding miRNA sponges of those validated circRNAs and integrated RBP informaiton from CircInteratome database. Then experimentally verified miRNA-disease associations and miRNA targets were united from several public databases. All those data formed a large regulation network to provide more information for users. Circ2Disease can be accessed at http://bioinformatics.zju.edu.cn/Circ2Disease/index.html

CircRNADisease is a database for experimentally supported circRNA dysregulations in human diseases. To obtain the high confident experimentally supported circRNA-disease associations, all circRNADisease entries were curated manually collected according to several steps as described below. First, the developer searched the PubMed database using the keywords 'circRNA disease', 'circular RNA disease', 'circRNA cancer' and 'circular RNA cancer' that had been recorded from the National Center for Biotechnology Information PubMed (NCBI-PubMed). Second, the developer manually extracted experimentally supported circRNA-disease associations. Different researchers were assigned to double-check all circRNA-disease pairs. Each entry in the circRNADisease includes detailed information on a circRNA-disease association, including circRNA (circBase ID, name, and synonym), disease basic information, the circRNA expression pattern , experimental detection techniques, circRNA associated genes and microRNAs, biological functions and molecular mechanisms of circRNA in disease, association with patients survival, a brief functional and expression description about circRNA, literature references and other annotation information. CircRNADisease can be accessed at http://cgga.org.cn:9091/circRNADisease/

TABLE S2. Details information of the four circRNA-disease datasets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | No. circRNAs | No. diseases | No. positive samples | No. unlabeled samples | Sparsity |
| circR2Disease | 561 | 100 | 607 | 55,493 | 1.09e-02 |
| CircAtlas | 708 | 117 | 775 | 82,061 | 0.94e-02 |
| Circ2Disease | 234 | 60 | 254 | 13,786 | 1.84e-02 |
| CircRNADisease | 286 | 48 | 304 | 13,424 | 2.26e-02 |

\* Sparsity indicates the ratio of positives to unlabeledsamples

## *Ablation studies*

To ensure the necessity of individual modules in the proposed model, we implemented the ablation studies. Specifically, we retained the basic framework of the model unchanged, eliminated the projector module, retained only the learner view, retained only the anchor view, and fused only the descriptors of the circRNA function and the disease MeSH, and thus constructed four different models and validated them on the CircR2Disease dataset, respectively, and the results are shown in Table S3. As can be seen from the table, the model after eliminating or replacing one of the modules in the proposed model achieves unsatisfactory results, averaging about 10 percentage points lower than GSLCDA in accuracy, F1, and MCC, with the non-projector model achieves better results only in sensitivity rate. The results of the ablation experiments indicate that each module in the proposed model has its own role and can contribute to the performance of the model.

TABLE S3. 5CV results obtained by different classifier models on CircR2Disease dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Acc. (%) | F1 (%) | Sen. (%) | MCC (%) |
| Non-projector model | 86.58±1.27 | 87.82±1.36 | 97.10±1.49 | 74.80±2.19 |
| Learner view model | 82.62±0.88 | 84.56±1.08 | 95.49±3.05 | 67.54±2.36 |
| Anchor view model | 80.89±1.38 | 82.77±2.17 | 92.37±4.01 | 63.46±3.40 |
| Different descriptors model | 83.44±2.65 | 83.90±3.00 | 86.77±1.91 | 67.03±5.11 |
| GSLCDA | **92.67±0.51** | **92.88±0.71** | **95.91±0.59** | **85.48±0.97** |